Soil Consistence and Structure as Predictors of Water Retention

W. J. Rawls and Ya. A. Pachepsky*

ABSTRACT

It is impractical to measure water retention for large-scale hydrologic, agronomic, and ecological applications or at the design stages of many projects; therefore, water retention estimates are often used. Field soil descriptions routinely include structure and consistence characterization. The objective of this work was to use the National Resource Conservation Service (NRCS) database to evaluate the potential for structural and consistence properties to serve as predictors of soil hydraulics properties. Total of ≈2140 samples were found that had (i) values of water contents at −33 kPa and −1500 kPa, (ii) structure characterized with grade, size, and shape, (iii) consistence characterized with dry and moist consistency, stickiness, and plasticity, and (iv) textural class determined in the field and from lab textural analysis. Because structural and consistence parameters were represented by categories rather than numbers, regression trees were used for recursive partitioning of the data sets into groups to decrease overall variability measured as the sum of squared errors within groups. Plasticity class, grade class, and dry consistency class were leading predictors of water retention at both −33 kPa and −1500 kPa matric potentials. The accuracy of estimates from structural and consistence parameters was lower than from textural classes. Using soil structural and consistence parameters along with textural classes provided a small, although significant improvement in accuracy of water retention estimates as compared with estimation from texture alone. Soil structural and consistence parameters can serve as predictors of soil water retention because those parameters reflect soil basic properties that affect soil hydraulic properties.

Knowing soil water retention is essential in many hydrologic, agronomic, and ecological applications. It is impractical to measure water retention for large-scale applications or at the design stages of many projects, and water retention estimates are often used.

Soil structural properties were shown to be an important factor of soil hydraulic properties. Void measurements and counts produce parameters of structure that help to estimate sol hydraulic conductivity. A detailed count of lengths and widths of voids allowed Anderson and Bouma (1973) to compute hydraulic conductivity of an argillic horizon of silt loam soil using a Kozeni-Karman equation for flows in slits. Pore size count was used by Rawls et al. (1993) to estimate the macropore K_{sat}. Lin et al. (1999a,b) presented an elaborated system of morphometric indices and showed that these indices, and not traditional texture, bulk density, and organic matter content, appeared to be the best predictors of parameters of macro- and micropore flow. Griffiths (1991) found ped size and number of biopores to be leading predictors for K_{sat} and hydraulic conductivity at −4 kPa. McKenzie et al. (1991) compared several sets of descriptive tables with mutually exclusive classes for hydraulic conductivity predictions.

Soil structural parameters were related to water retention. Williams et al. (1983) found that the presence of pedality, particle-size distribution, and grade of structure were the soil properties most consistently associated with similar moisture characteristics. Williams et al. (1992) and Danalatos et al. (1994) suggested having different equations to estimate soil water retention for weakly structured and well-structured soil horizons. Hall et al. (1977) and McKeague (1987) used regional databases from Canada and England to select soil structural parameters suitable to estimate air-filled porosity and available water capacity.

Soil consistence was also referred to with respect to soil hydraulic properties. Bicki et al. (1988) found that Mollisols had higher percolation rates and lower coefficients of variation than Alfisols within a biosequence, and attributed this difference, in part, to stronger subsoil consistence in Alfisols. Vepraskas et al. (1996) observed that parallel changes in saturated hydraulic conductivity and in stickiness and plasticity occur in transitional horizons separating soil and saprolite. Studying a variety of soils, Voronin (1990) found a linear relationship between gravimetric water contents at plasticity limit and soil water potential. Data on consistence in terms of penetration resistance were recently found to be useful predictors of soil water retention (Pachepsky et al., 1998; Gimenez et al., 2001). It was recognized that databases for predicting soil water retention have to be expanded to include soil structure and soil consistence parameters (Rawls et al., 1991).

Soil structure can be characterized with either categorical or numerical variables. Categorical characterization consists in setting classes or categories, like weak, moderate, and strong for the grade, and recording the class or category for each soil sample. Numerical characterization presumes the ability to measure the structural variable and to have a range of continuous values to characterize soil samples, like it is done for bulk density or penetration resistance. When soil structural properties of shape, size, and grade are presented as categorical rather than numerical variables, data about soil structure cannot be directly used in statistical regressions or neural networks to estimate water retention from other soil properties. The same is true for soil consistence properties if they are characterized in categorical rather than numerical terms. It was shown that the categorical variables can be used to set boundaries to partition soils into several hydraulic conductivity classes (Wang et al.,

W.J. Rawls, USDA-ARS Hydrology and Remote Sensing Lab., Bldg. 007, Rm. 104, BARC-WEST, Beltsville, MD 20705. Ya.A. Pachepsky, USDA-ARS Animal Waste Pathogen Lab., Bldg. 173, Rm. 203, BARC-EAST, Beltsville, MD 20705. Received 27 July 2001. *Corresponding author (ypachepsky@anri.barc.usda.gov).

Published in Soil Sci. Soc. Am. J. 66:1115–1126 (2002).

Abbreviations: θ, water content at the matric potential of interest; M_{group}, minimum number of samples in a group after partitioning; M_{split}, minimum number of samples before a partitioning; NRCS, National Resource Conservation Service; RMSE, root mean squared error.

1985). However, a procedure to select an optimum partitioning for hydraulic conductivity or water retention was not suggested. Recently, regression trees were recognized as a suitable statistical technique for using categorical variables as predictors (Clark and Pregibon, 1992). Regression trees were successfully used to explore databases in natural sciences (Michaelson et al., 1994; Fielding, 1999), and, in particular, in soil science (McKenzie and Jacquier, 1997; Anderson et al., 1999). Optimum partitioning of databases with regression trees was used to find both the best predictors and best grouping of samples.

Field soil descriptions routinely include structure and consistence characterization. Therefore, it can be beneficial to know whether and to what extent structural and consistence properties may serve as predictors of soil hydraulic properties that are more difficult to measure. Those questions can be addressed with data from the NRCS database (Soil Survey Staff, 1997), as it contains a large volume of coupled data on soil structure, soil consistence, and soil water retention at −33 and −1500 kPa. Regression trees are a tool of choice because the NRCS data sets include categorical rather than numerical variables.

The objective of this work was to use the NRCS database to evaluate the potential for structural and consistence properties to serve as predictors of soil hydraulics properties.

MATERIALS AND METHODS

Soil Database

The National Soil Characterization Database was screened to select soil samples that had (i) values of water contents at −33 kPa and −1500 kPa on clods and bulk densities at −33 kPa and air dried soil, (ii) structure characterized with grade, size, and shape, (iii) consistence characterized with dry and moist consistency, stickiness, and plasticity, and (iv) textural class determined in the field and from lab textural analysis, all measured and described in the same pedon. A total of 2142 and 2137 samples were found for −33 kPa and −1500 kPa matric potentials, respectively. Thirty percent of all samples in that data set belonged to pedons that did not have a taxonomic family phrase. Mollisols, Aridisols, Alfisols, and Entisols were the most numerous among soils with known taxonomy in the data set, and constituted 24, 14, 11, and 6%, respectively. About half of all samples came from California, Colorado, Idaho, Kansas, New Mexico, Texas, and Washington.

Figure 1 shows distributions of structural and consistence properties among samples in the data set. The weak and moderate grades are the most common in the data set, whereas samples with the strong grade constitute only ≈10%. Medium and fine sizes dominate in the data set. Angular blocky, blocky, and subangular blocky shapes were by far overrepresented in the dataset. No columnar, massive, single grain shape was found, and only 14 samples had the wedge shape. The moist consistence in all but 17 samples was friable or very friable. The dry consistency was hard to various degrees in ≈90% of samples. Stickiness was well distributed between various categories, from nonsticky to very sticky. Plasticity was also well distributed, and the slightly and moderately plastic samples were more numerous than samples that were very plastic or nonplastic.

The major field-determined textural class in the data set was silt loam found in ≈24% of all samples (Table 1). Sandy loam, loam, clay, and silty clay loam were represented with 15, 12, 12, and 10% of all samples, respectively. Silt and sandy clay were each represented with <0.5% of all samples, while sands and loamy sands were each ≈3% of all samples. Values of volumetric water contents at −33 kPa, θ_{33}, and −1500 kPa, θ_{1500} (θ is the water content at the matric potential of interest), were obtained as products of gravimetric water contents on corresponding bulk density.

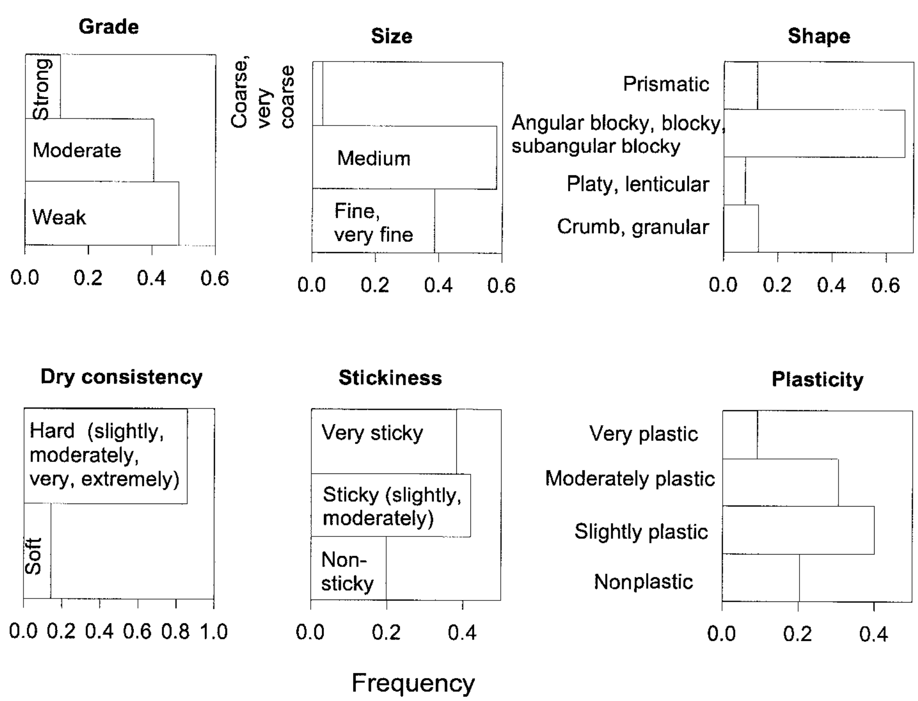

Fig. 1. Distributions of structural and consistence parameters among the samples in the data set of this work.

Table 1. Statistics of water retention (%, by vol.) at −33 and −1500 kPa matrix potential by textural classes defined from field judgement and laboratory determination of texture.

	Field-judged textural class					Textural class from laboratory measurements				
		−33 kPa		−1500 kPa			−33 kPa		−1500 kPa	
Textural class	*n*	Mean	SD	Mean	SD	*n*	Mean	SD	Mean	SD
Sand	69	12.7	7.9	4.1	2.4	12	10.9	9.5	4.1	4.8
Loamy sand	66	16.3	8.3	6.6	2.5	63	15.0	7.4	5.6	3.0
Sandy loam	319	22.5	8.4	10.0	4.1	333	20.9	8.2	8.8	3.8
Loam	257	28.9	6.9	14.4	5.1	213	29.9	8.6	14.3	5.1
Silt loam	510	33.1	6.3	13.7	4.5	479	32.8	6.7	13.8	4.9
Silt	6	34.3	5.2	9.1	5.2	1	34.0	NA	7.7	NA
Sandy clay loam	76	28.2	6.3	16.8	4.9	105	25.0	6.7	13.6	4.3
Clay loam	200	34.9	5.5	20.3	4.1	203	33.6	7.0	18.9	5.1
Silty clay loam	221	36.5	5.7	20.7	4.1	288	36.9	5.9	21.0	5.4
Sandy clay	8	26.4	4.0	19.2	4.7	10	27.7	5.2	17.6	3.2
Silty clay	149	40.2	4.9	26.4	4.4	132	39.1	5.2	26.0	4.8
Clay	261	41.4	6.5	28.5	5.1	303	40.1	6.3	27.1	5.3

Regression Tree Modeling

Regression tree modeling is an exploratory technique based on uncovering structure in data (Clark and Pregibon, 1992). The resulting model partitions data first into two groups, then into four groups, and so on, providing groups as homogeneous as possible at each of the levels of partitioning. Each partitioning can be viewed as a branching, and the final fit of model to data looks like a tree with two branches originating in each node. Regression trees first became quite popular in environmental sciences (Lees and Ritman, 1991; Baker, 1993), and were later used in studies on land quality assessment and soil properties estimation (Van Lanen et al., 1992; McKenzie and Jacqier, 1997; McKenzie and Ryan, 1999). Regression trees can use both categorical and numerical variables as predictors (Breiman et al., 1993). Regression trees can be developed with the software SPLUS (MathSoft, 1999) that has been used in this work, and also with the SAS software.

The regression tree algorithm works as follows. Suppose that a database is organized as a table with columns x_1, x_2, x_3,..., x_N representing predictor variables and the column y representing the response variable. The minimum number of samples, or database lines, before a partitioning, M_{split}, and the minimum number of samples in a group after partitioning, M_{group}, has to be set first. Then all possible partitions are formed for each of predictor variables.

The method of forming a partition depends on the type of the variable. If the variable is numerical, then the whole database table is sorted by the column of this variable and then split up into two parts, each having the number of samples greater than M_{group}. An example of such partitioning is shown in Table 2, where raw data are shown in Columns 1 to 4: variable x_1 is numerical; variable x_2 is categorical with three possible levels *a*, *b*, and *c*; and *y* is the variable of interest to be predicted. Columns 4 to 8 show the same database sorted by the numerical variable x_1. Let the value of M_{group} be equal to 3. Then, possible partitions of the database by the variable x_1 are:

(1) either $x_1 < 1.6$, Samples 1, 3, 6, or $x_1 > 1.6$, Samples 11, 5, 2, 9, 8, 10, 7, 4;
(2) either $x_1 < 2.15$, Samples 1, 3, 6, 11, or $x_1 > 2.15$, Samples 5, 2, 9, 8, 10, 7, 4;
......
(7) either $x_1 < 3.1$, Samples 1, 3, 6, 11, 5, 2, 9, 8, or $x_1 > 3.1$, Samples 10, 7, 4.

If the variable is categorical, then the partitions are formed after dividing the database into subsets having the same value of this variable and splitting it by all possible combinations of subsets. In the example of Table 2, the subsets are shown in columns 9 through 13. Then possible partitions are

(1) a|bc, i.e., either x_2 = a, Samples 5, 6, 9, or x_2 = b or x_2 = c, Samples 1, 3, 4, 10, 11, 2, 7, 8.
(2) ab|c, i.e., either (x_2 = a or x_2 = b), Samples 5, 6, 9, 1, 3, 4, 10, 11, or x_2 = c, Samples 2, 7, 8.
(3) ac|b, i.e., either (x_2 = a or x_2 = c), Samples 5, 6, 9, 2, 7, 8, or x_2 = b, Samples 1, 3, 4, 10, 11.

Table 2. Synthetic database to explain the regression tree algorithm.

Raw data				Data sorted by the variable x_1					Data sorted by the variable x_2			
Sample	x_1†	x_2‡	y§	Sample	x_1	x_2	y	ΔD¶	Sample	x_1	x_2	y
1	0.8	b	7.9	1	0.8	b	7.9	ND#	5	2.4	a	3.2
2	2.7	c	4.8	6	1.1	a	3.5	ND	6	1.1	a	3.5
3	1.3	b	6.1	3	1.3	b	6.1	ND	9	2.8	a	3
4	4.2	b	2.5	11	1.9	b	5.5	14.1	1	0.8	b	7.9
5	2.4	a	3.2	5	2.4	a	3.2	19.7	3	1.3	b	6.1
6	1.1	a	3.5	2	2.7	c	4.8	14.5	4	4.2	b	2.5
7	3.6	c	0.6	9	2.8	a	3	18.5	10	3.3	b	1.7
8	2.9	c	5	8	2.9	c	5	14.7	11	1.9	b	5.5
9	2.8	a	3	10	3.3	b	1.7	23.4	2	2.7	c	4.8
10	3.3	b	1.7	7	3.6	c	0.6	ND	7	3.6	c	0.6
11	1.9	b	5.5	4	4.2	b	2.5	ND	8	2.9	c	5

† x_1 variable is numerical.
‡ x_2 variable is categorical to three levels.
§ y is the variable of interest to be predicted.
¶ ΔD = change in deviance.
# ND = cannot be defined.

The total possible number of partitions with a categorical variable is $2^{k-1} - 1$, where k is the number of levels. For example, for a variable having four levels a, b, c, d, there are $2^{4-1} - 1 = 7$ possible partitions (a|bcd, ab|cd, abc|d, ac|bd, ad|bc, abd|c, acd|b).

The first subset of samples in a partition is called left branch, and the second is called right branch. For example, in the partition (1) of the database in Table 2 by the variable x_1, the left branch will consist of samples where $x_1 < 1.6$ (Samples 1, 3, 6), and the right branch consists of samples where $x_1 > 1.6$ (Samples 11, 5, 2, 9, 8, 10, 7, 4).

All partitions by all variables are compared with the reduction in nonhomogeneity that they provide. The nonhomogeneity in a group of samples is measured by computing deviances, which are defined for a group of observed values y as

$$D = \sum_{i} (y_i - \bar{y})^2$$

Here, $\bar{y}$ is the mean value across all observations y_i. Each partition generates left $D_L = \sum_{L} (y_i - \bar{y})^2$ and right $D_R = \sum_{R} (y_i - \bar{y})^2$ deviance values where subscripts L and R indicate collections of numbers of samples in branches in a partition. The partition that maximizes the change in deviance

$$\Delta D = D - D_L - D_R$$

is the partition to chose. For the example of Table 2, values of ΔD for the partitioning by the variable x_1 are shown in the column 9. The largest change in deviance, $\Delta D = 23.4$, is achieved after partitioning 'either $x_1 < 3.1$ or $x_1 > 3.1$'. Values of ΔD obtained after partitioning by the categorical variable x_2 are $\Delta D = 2.3$ for a|bc, $\Delta D = 1.1$ for ab|c, and $\Delta D = 5.6$ for ac|b. Therefore, the partitioning 'either $x_1 < 3.1$ or $x_1 > 3.1$' is the one to choose in this example.

Each branch obtained after partitioning is partitioned again in accordance with the limitations imposed by values of M_{split} and M_{group}. In the example of Table 2, the group of samples where $x_1 > 3.1$ cannot be partitioned anymore, but the group with $x_1 < 3.1$ can. Attempts to partition the latter group by the variable x_1 lead to the best values of $\Delta D = 6.1$, whereas partitioning 'either $x_2 = a$ or ($x_2 = b$ or $x_2 = c$)' gives the value of $\Delta D = 12.9$, and becomes the partitioning to choose. Further partition is not possible because the minimum size of a group is achieved, and this node of the tree is a terminal node. The final regression tree for this example is shown in Fig. 2. Here the number of a terminal node is shown in brackets, and the average value of the variable y for this node, standard deviations of the variable y in parentheses, and the count of samples pertaining to this node are shown beneath the terminal node numbers. The average value is the value predicted for the whole group of samples forming the terminal node.

In the limit, the recursive partitioning of a large database may produce a tree with a very large number of terminal nodes. There is a chance that the predictive ability of such large tree will be limited, because the later branches will show intricacies of small groups of samples specific for the database. To avoid such *overfitting*, a tree has to be pruned to be useful for predictions. The regression tree methodology has variations regarding tree pruning (Bell, 1999). To snip off the least important partitions, the software SPLUS (MathSoft, 1999) that has been used in this work, applies the cost-complexity measure D_K:

$$D_k(T) = D(T) + K \times N_{TN}$$

Here $D(T)$ is the deviance of the subtree (T), N_{TN} is the number of terminal nodes, and K is the cost-complexity parameter.

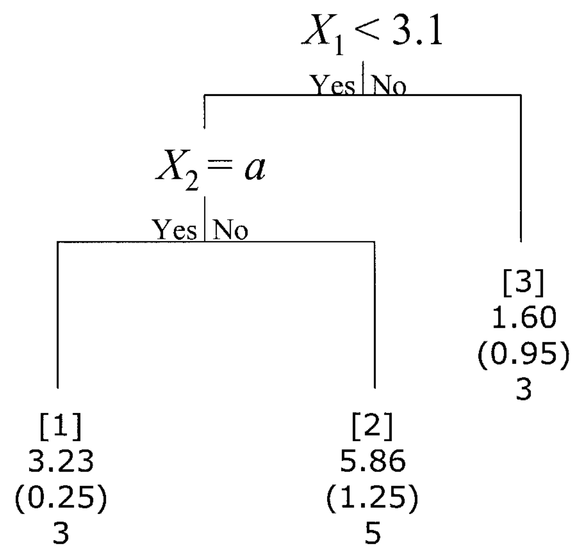

Fig. 2. Regression tree for the test data set in Table 1. The node number in brackets, the average value of the dependent variable for the group, the standard deviation of the dependent variable within groups in parentheses, and the number of samples in the group are shown beneath terminal nodes.

To decide on the value of K for this study, we applied the jackknife cross-validation (Good, 1999). The original database of 2142 samples was 10 times randomly divided into development and testing subsets in 9:1 proportion. For each such division, regression trees were obtained for the development subsets and then applied to both development and testing subsets. Accuracy of the trees was estimated with the data that had been used to obtain the regression tree (development data set), and with data that had not been in the tree construction (testing datasets). The accuracy of water retention estimates was characterized with root mean squared error (RMSE) for both development and testing subsets. The RMSE was computed as $\text{RMSE} = \sqrt{\sum_{i} (\theta_{i,est} - \theta_{i,meas})^2 / N_d}$, where subscripts "est" and "meas" denote predictions from regression trees and measured data, respectively; N_d is the number of samples in the subset minus the number of partitions, summation is over all data in the subset. Average values of RMSE for each value of K from those computations are shown in Fig. 3 for water retention at -33 kPa. As the parameter K decreases, the accuracy of estimates in both development and testing improves (Fig. 3a), and the number of terminal nodes increases (Fig. 3b). The accuracy of estimates in the testing data sets stabilizes somewhere between $K = 100$ and $K = 400$, and remains approximately constant when values of K further decrease. The variability in RMSE in testing data sets is such that the differences between RMSE at $K = 400$ and $K = 100$ are not statistically significant (Fig. 3a). Because the average number of terminal nodes is significantly smaller for $K = 400$ than for $K = 100$ (Fig. 3b), we preferred using the larger value of K to have fewer groups to simplify interpretation of results of partitioning in those groups. Similar results were obtained for water retention at -1500 kPa (data not shown). The value of $K = 400$ was used for both -33 and -1500 kPa. Values of $M_{split} = 10$ and $M_{group} = 5$ were used in all computations. Hypotheses about the significance of differences between average values were tested with the t-test at the significance level of 0.05.

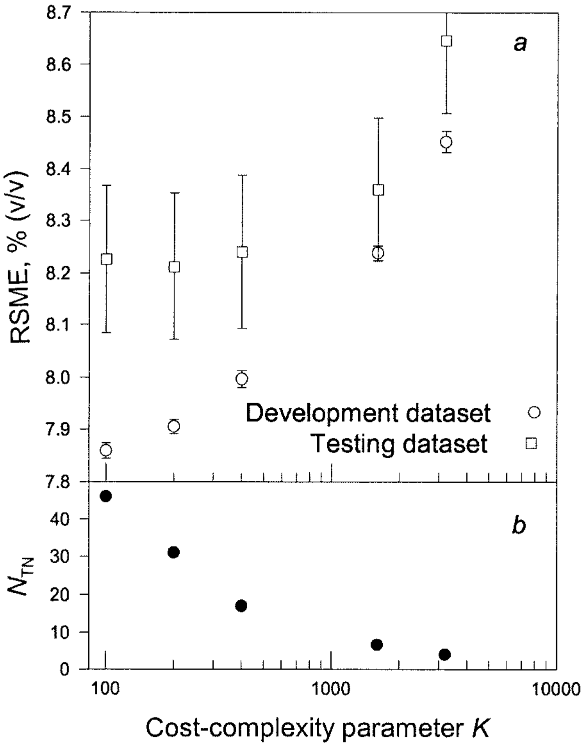

Fig. 3. Effect of the size-complexity parameters on the accuracy, reliability, and number of terminal nodes N_{TN} of the regression tree. RMSE, root mean squared error of the volumetric water content at −33 kPa.

RESULTS

Water Retention at −33 kPa

The regression tree obtained for water content at −33 kPa from the whole data set is shown in Fig. 4. The first partition of samples into two most homogeneous groups occurs by the plasticity, so that nonplastic samples form one large group, and plastic samples constitute another one. Among nonplastic samples, samples with weak and moderate grade form one large group, and samples with strong grade form a separate Terminal Node [9] with the average θ_{33} value that is significantly larger than in most of other nonplastic samples. The stickiness is the next important partitioning factor in weak or moderate grade samples. Nonsticky samples form one group, and sticky to various extent samples are assembled in another group. Nonsticky samples are further partitioned by their grade; the moderate-grade samples form a Terminal Node [5], whereas the weak-grade samples are partitioned by the shape of structural units. The samples with blocky or prismatic shape of structural units have smaller average water retention than samples with other shape of units combined in Terminal Node [4]. Among samples with blocky or prismatic shape of structural units, samples with soft dry consistency have smaller average water retention (Nodes [1] and [2]) as compared with samples with hard dry consistency (Node [3]). Nonplastic samples with structure showing weak or moderate grade, blocky shape, and soft consistency are finally partitioned by the size of structural units (Nodes [1] and [2]).

Nonplastic samples of weak or moderate grade that are sticky to some extent are partitioned by the size of structural units. Medium or coarse size leads to the smaller average water retention (Node [6]) as compared with samples having the fine size of structural units at Terminal Nodes [7] and [8]. The two latter nodes are distinguished by the grade of the samples, so that the weak grade results in larger average water content at −33 kPa (Node [8]).

The samples that are slightly plastic are first partitioned in the same way as nonplastic samples, that is, by the grade and then by stickiness. Slightly plastic samples of weak grade have average water retention increasing as the stickiness increases (Nodes [9], [10], and [11]). Among slightly plastic samples with moderate or strong grade, stronger grade leads to the larger θ_{33}, as shown in Nodes [12] and [13].

Moderately plastic samples are partitioned by the dry consistency, so that soft dry consistency resulted in smaller average θ_{33} (Node [14]), as compared with the hard dry consistency (Node [15]). Finally, very plastic samples could be partitioned by the shape of structural units, and the blocky shape resulted in smaller average water content at −33 kPa (Node [16]), as compared with other shapes of structural units (Node [17]).

The main decrease in the deviance occurs during the first two or three partitions. In plastic samples, partitioning by the degree of plasticity decreases the deviances more than partitioning by grade and stickiness does for nonplastic samples (data not shown). The more plastic the samples are, the smaller the variability is in θ_{33} within groups of samples, as characterized by standard deviations. The count of samples in a group does not seem to affect the variability within the group.

Water Retention at −1500 kPa

The regression tree obtained for water retention at −1500 kPa from the whole data set is shown in Fig. 5. Plasticity is the leading split variable for the θ_{1500}, as it was for θ_{33}. First, two large groups are formed by separating nonplastic and slightly plastic on one hand, and moderately plastic and very plastic on another hand. Non- or slightly plastic samples that are not sticky form three groups in which the average water retention at −1500 kPa increases as the grade becomes stronger (Nodes [1], [2], and [3]). Non- or slightly plastic samples with various degrees of stickiness are partitioned by grade of structural units. Where grade is weak or moderate, a further partitioning occurs according to the dry consistency. The soft dry consistency leads to smaller water retention at −1500 kPa (Node [4]), as compared with the hard dry consistency where the largest θ_{1500} corresponds to very sticky samples (Node [8]). Where samples are non- or slightly plastic, have weak or moderate grade of textural units, and are moderately sticky, the water content at −1500 kPa increases as the shape of textural units changes from 'platy or lenticular or prismatic' to 'blocky or angular blocky' to 'crumb or granular', as shown in Nodes [5], [6], and [7]). Non- or slightly plastic samples with various degrees of stickiness and

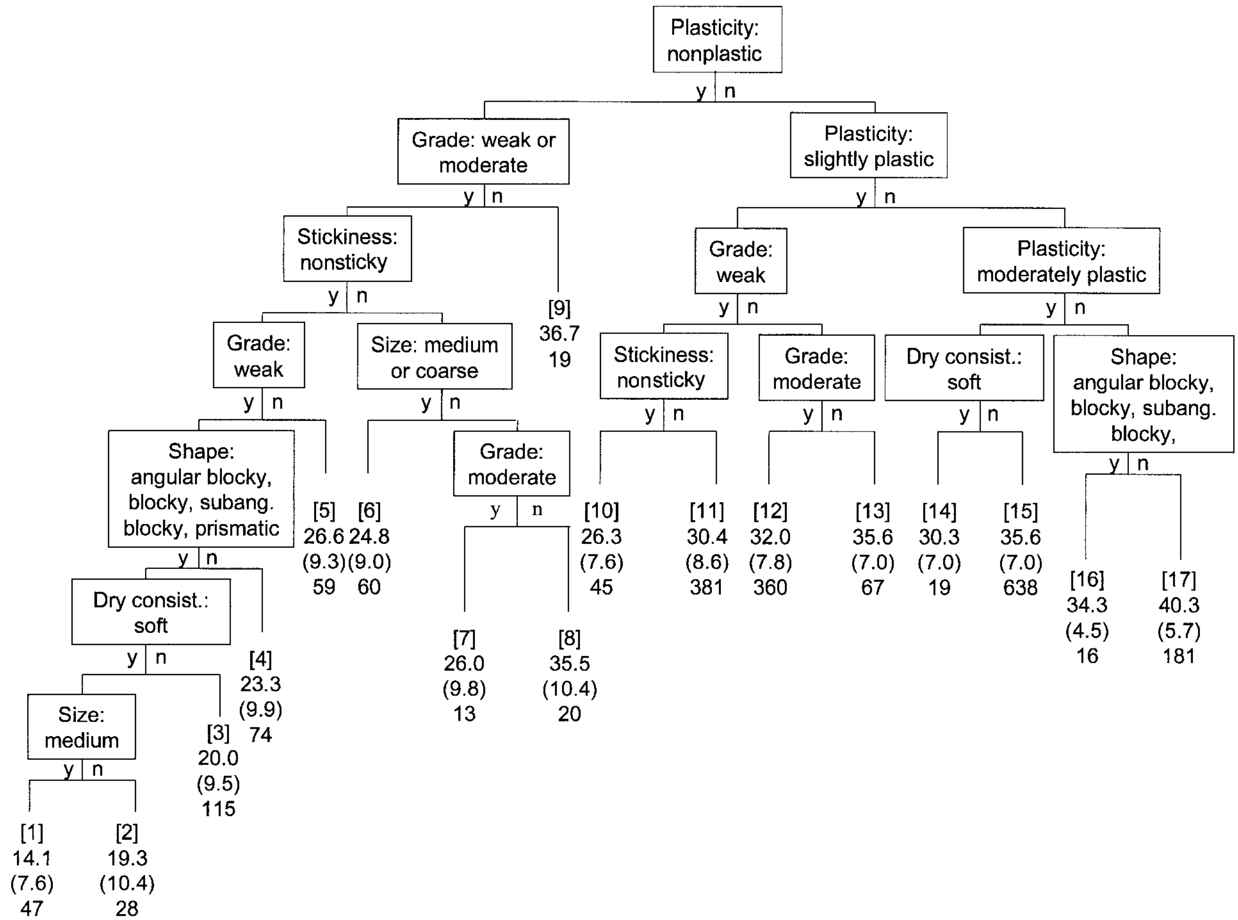

Fig. 4. Regression tree to estimate water retention at −33 kPa from structural and consistence parameters; y, yes and n, no answers to the parameter definition in the box above. The node number in brackets, the average value of the volumetric water content at −33 kPa for the group, the standard deviation of the water content within groups in parentheses, and the count of samples in the group are shown beneath terminal nodes.

strong grade are partitioned by the shape of structural units so that platy, lenticular, or prismatic units correspond to smaller average water content at −1500 kPa (Node [9]), as compared with other shapes (Node [10]).

Moderately plastic samples are split into subgroups by their dry consistency (Nodes [11] and [12]) so that the soft dry consistency means smaller water retention at −1500 kPa. Very plastic soils form a single group with the largest average θ_{1500} (Node [13]).

Using stickiness along with plasticity brings a substantial decrease in deviance for non- and slightly plastic samples. The main decrease in the deviance occurs during the first two or three partitions. No relationship between the degree of plasticity and within-group variability can be seen in Fig. 5. The count of samples in a group does not affect variability in θ_{1500} within the group.

Combining Textural Classes with Structural and Consistence Parameters to Estimate Water Retention

Accuracy of water retention estimates from field structural and consistence parameters is compared in Table 3 with accuracy of estimates from textural class determined from laboratory analysis. The estimates from laboratory textural classes are about 20 to 40% more accurate. The root-mean square errors are 8 and 6.7% (by vol.) for −33 kPa and 5.6 and 4.4% (by vol.) for −1500 kPa from structural and consistence parameters and from laboratory textural classes, respectively. The root-mean square errors of estimates from field textural classes are ≈7 and 4.8% (by vol.) for −33 and −1500 kPa, respectively. Variability of water retention at both −33 and −1500 kPa within the groups found with regression trees (Fig. 4 and 5) is comparable with the variability within textural classes (Table 1). Differences between average values in the groups are not statistically significant for grouping either with regression trees or by field-judged or lab-determined textural classes.

Adding structural and consistence parameters to laboratory textural classes brings a small, albeit significant, increase in accuracy of estimates (Table 4). The regression trees are shown in Fig. 6. In coarse textural soils, plasticity and stickiness follow textural class in partitioning data at −33 kPa (Fig. 6a). Dry consistency and shape are most useful to group data on fine textural soil at this matrix potential. Plasticity, stickiness, and grade are helpful to partition water retention at −1500 kPa only in loams, silt loams, and sandy clay loams (Fig. 6b).

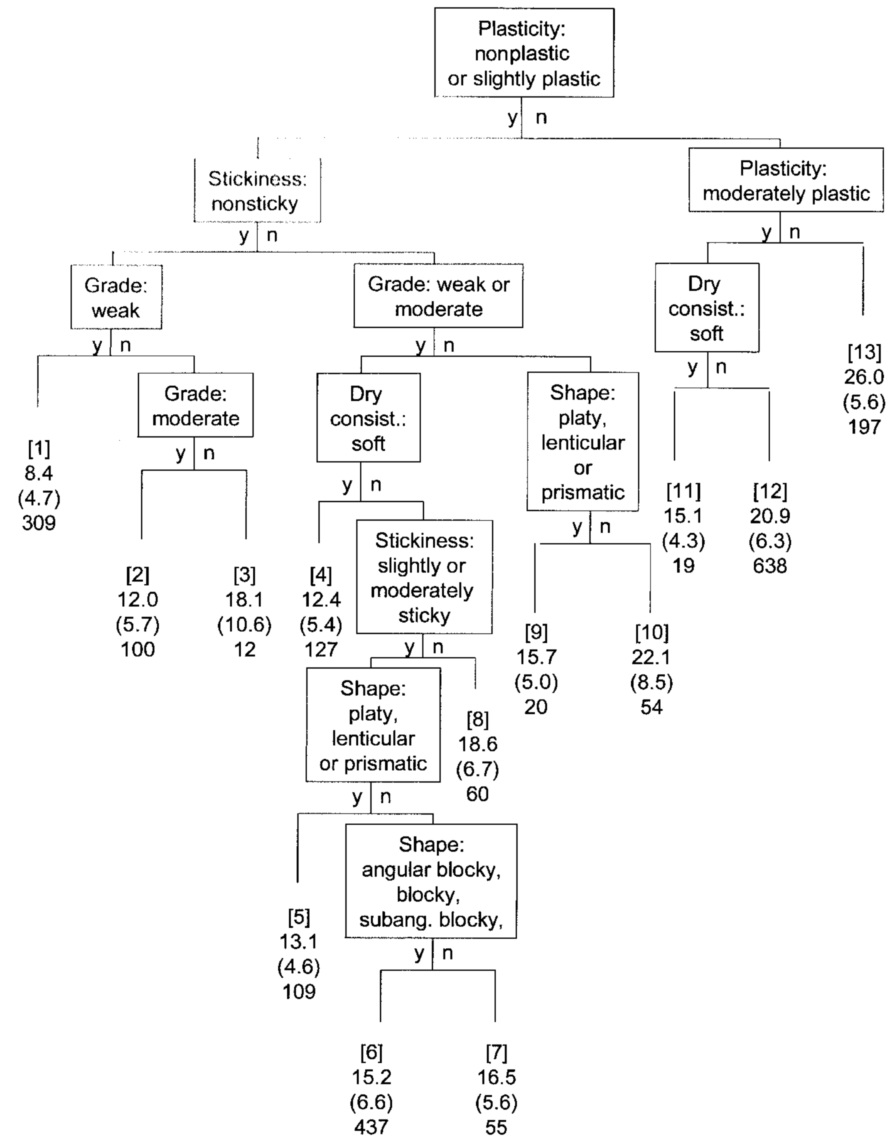

Fig. 5. Regression tree to estimate water retention at −1500 kPa from structural and consistence parameters; y, yes and n, no answers to the parameter definition in the box above. The node number in brackets, the average value of the volumetric water content at −1500 kPa for the group, the standard deviation of the water content within groups in parentheses, and the count of samples in the group are shown beneath terminal nodes.

DISCUSSION

Field-determined structural and consistence categorical parameters provide enough information to be used in regression trees for partition soil samples by their water retention. All those parameters are defined only by two to four broad classes, and are to some extent observer-specific (Nettleton et al., 1969; Post et al., 1986). Nevertheless, the accuracy of the water retention estimates lies within the range of accuracy achieved using much more information from laboratory analyses. Schaap and Leij (1998) reported RMSE values of 10–

Table 3. Probability distribution of errors (%, by vol.) in water retention estimates from field structural and consistence parameters and from textural class determined from laboratory analysis.

Probability	Structural and consistence parameters		Textural class determined in laboratory	
%	−33 kPa	−1500 kPa	−33 kPa	−1500 kPa
10	−10.0	−6.7	−7.9	−5.1
25	−5.3	−4.1	−4.4	−3.0
50	−0.3	−0.9	−0.5	−0.4
75	5.0	3.1	3.6	2.3
90	9.9	7.8	8.0	5.6

Table 4. Root mean square errors (%, by vol.) of water retention with various sets of categorical soil basic parameters.

Data set	−33 kPa	−1500 kPa
Lab texture class	6.7	4.4
Lab texture class + structure	6.6	4.4
Lab texture class + consistence	6.5	4.3
Lab texture class + structure + consistence	6.4	4.3
Field texture class	7.0	4.8
Field texture class + structure	6.9	4.8
Field texture class + consistence	6.9	4.8
Field texture class + structure + consistence	6.8	4.8
Structure + consistence	8.0	5.6

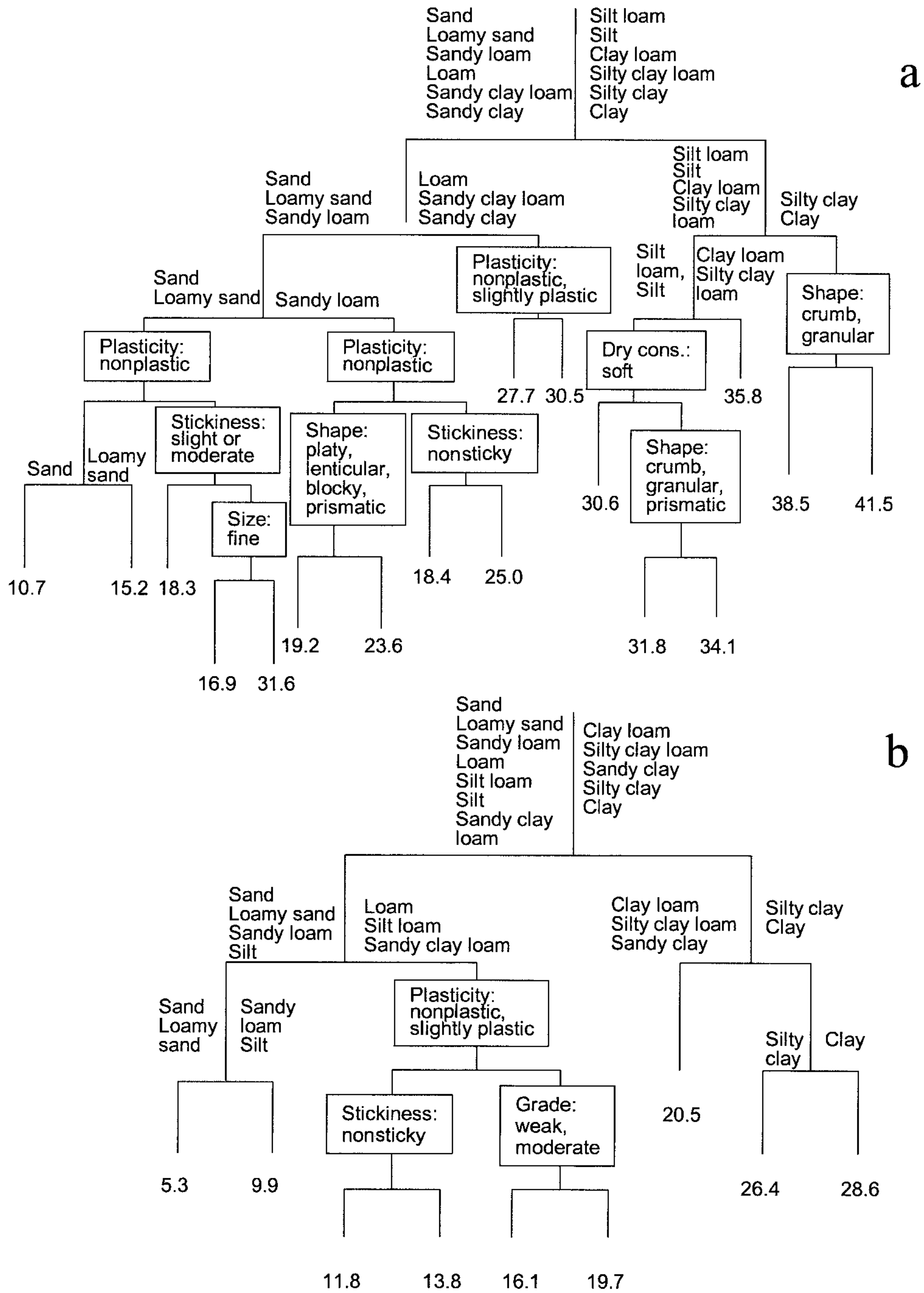

Fig. 6. Regression trees to estimate water retention from laboratory textural class and structural and textural parameters at (a) −33 kPa and (b) −1500 kPa. Left (right) branches lead to samples where the categorical variables have (have not) values shown in the box above splits; the average value of the volumetric water content at −33 kPa for the group is shown beneath terminal nodes.

12% (by vol.) for estimates of water retention from three large interregional databases using artificial neural network with sand, silt, and clay contents and bulk density from laboratory analyses as input variables. Leenhardt (1995) estimated water retention in a small regional database from laboratory data on clay content with RMSE ≈7% (by vol.). This comparison indicates that structure and consistence determined in the field contain substantial information about soil properties relevant to soil water retention.

It seems to be appropriate to discuss trends in changes of group average values as related to soil properties. Soil plasticity is a leading variable for partitioning soil samples by their water retention. The ability of soils to exhibit plastic behavior has long been thought to be related to soil clay content, and soil organic matter is also a recognized factor increasing plasticity (Horn and Baumgartl, 1999). Figure 7a shows the distribution of plasticity classes within textural groups in our data set. As expected, sands and loamy sands are mostly nonplastic. However, a substantial amount of sandy loams, loams, silt loams, silts, and sandy clay loams are nonplas-

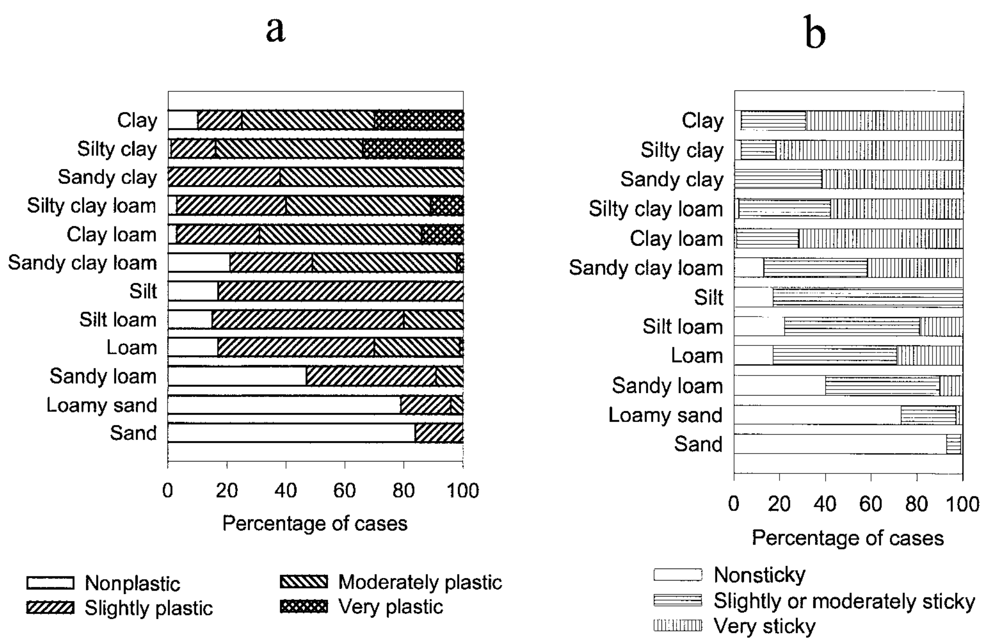

Fig. 7. Distribution of plasticity classes and stickiness classes among textural classes.

tic, and even 10% of clay samples are nonplastic. In nonplastic soils, the next important parameter for grouping is grade (Fig. 4). Nonplastic soils with strong grade, that is, with distinct structural units, probably have large pore space available for water storage at −33 kPa, that is, close to field capacity. This may explain the large average value of θ_{33} at Terminal Node [8] in Fig. 4. Other partitions by grade also lead to increase of the average θ_{33}, as grade changes from weak to moderate to strong (i.e., in Fig. 4, Nodes [1]–[5] as compared with Nodes [6]–[8]; Nodes [10] and [11] as compared with Nodes [12] and [13], and Nodes [7] and [12] as compared with Nodes [8] and [13], respectively). Interestingly, an increase in grade leads also to the increase in water retention at −1500 kPa (see Nodes [1]–[3] and Nodes [4]–[7] in comparison with Nodes [9] and 10 in Fig. 5). This may mean that the visible well-pronounced grade remains well expressed at finer scales where water retention at −1500 kPa takes place, or that the distribution of a structural units' properties reflects the pore size distribution at finer scales (Filgueira et al., 1999).

The observed effect of grade on the average θ_{33} is similar to the one reported for water retention at −10 kPa θ_{10} by several authors. Bouma (1992) observed differences in water retention between weak and strong grade in arable and grassed Haplaquent, respectively, both having subangular blocky structure. The average water retention at −10 kPa was larger in samples with strong grade (Bouma, 1992), although this difference was not statistically significant. Soil with a weaker grade also had smaller water retention at −10 kPa in the study of Bouma and Anderson (1973), who compared water retention of two fine silty mesic Argiudolls, both having medium prizmatic parting to subangular blocky structural units. Yet another insight in the importance of the grade give data from the study of Shaw et al. (1997), where pore size distributions have been compared for B_{tv} and B_t horizons in 18 pedons of fine loamy, siliceous, thermic Kandiudults with various contents of plinthite. Image analysis showed much larger percentage of pores with the equivalent diameter between 0.05 and 0.005 cm in horizons with weak grade as compared with horizons with the moderate grade. That range of equivalent diameters corresponds to the range of matric potentials between 0.6 and 6 kPa, which means that soil in the horizon with a weak grade loses much more water as the suction is applied as compared with the soil in horizons with the moderate grade. Southard and Buol (1988) observed that in Ultisols that they had studied, grade of blocky structure gradually became stronger with depth, whereas the amount of pores emptying at −10 kPa decreased with depth. This meant an increase in water retention since bulk density did not show depth-related trends. Grade appears to be a relatively strong predictor of water retention, and grade-related parameters of aggregate size distribution have been included in soil tilth index (Singh et al., 1992) and in the model of soil water retention with explicit formulation of structure effects (Nimmo, 1997).

An increase in stickiness for the same plasticity class necessarily leads to the increase in water retention both at −33 and −1500 kPa. Various classes of stickiness can be found in the same textural class (Fig. 7b), although a trend of increase in stickiness with the decrease of sand content can be traced in this figure. As opposed to plasticity, an increase in organic matter content tends to decrease stickiness (Domzal, 1970; Chancelor, 1994). Different types of clays exhibit widely different stickiness characteristics (Chancelor, 1994). For example, montmorillonite clay is three times more sticky than kaolinite clay. Clay mineralogy and organic matter content are important factors of soil water retention (Rawls et al., 1991). Stickiness is affected with those factors, too, and this is a probable reason for stickiness being the second important consistence property to be related to soil water retention, as shown in Fig. 4 and 5.

Size of structural units is included in the partitioning variables only for nonplastic soils for θ_{33}. Soils of weak grade have larger average values of water retention at −33 kPa if their structural units are finer (Nodes [1], [2], and Nodes [6]–[8] in Fig. 4). The absence of large structural units and weak grade may mean absence of large pores and a wide pore-size distribution that should provide relatively large water retention near field capacity.

Shape of structural units enters the regression at a relatively late stage of partitioning if textural classes are not used. In nonplastic soils, blocky shape of structural units corresponds to the lower average water retention at −33 kPa, as compared with other shapes (Nodes [1], [2], and [3] as compared with Node [4] in Fig. 4). In very plastic soils, the effect is quite the opposite (Nodes [16] and [17]). We hypothesized that, in the latter case, blocky shape of soil structural units reflects the presence of minerals with mobile lattice that enhance water retention of soils. Another reason may be that soil textural differences create differences in role of shape of structural units in water retention. An indication can be found in Fig. 4, where crumb or granular shape results in the smallest water retention in sandy loams and in the largest water retention in silt loams, silt, clay loams, silty clay loams, and clays at −33 kPa. Nonsticky soils with platy-, lenticular-, or prismatic-shaped structural units have lower average water retention at −1500 kPa, and this trend probably has to be attributed to specifics of soil composition that we have not been able to discern.

Soil dry consistency expresses a soil resistance to rupturing or deformation, and is usually interpreted with reference to cohesion between soil particles. Whereas plasticity is the property which allows deformation without cracking, cohesion is a possession of a shear strength which allows the soil to maintain its shape under load even when it is not confined (Carter and Bentley, 1991). Although dry consistency generally increases with the increase of clay content, it is affected both with contents of other textural fractions (Ibanga et al., 1980) and with intricacies of the fine fraction composition. Carter and Bentley (1991) state that whereas plasticity is produced by the electrochemical nature of the clay particles, cohesion occurs as a result of their very small size. Nevertheless, clay mineralogy is an important factor of soil strength (Horn, 1993). In loess soils that have been air-dried, carbonates create a rigid frame that increases resistance to rupture (Lysenko, 1972). Carbonates and other salts precipitating from soil solutions produce the same effect in clay soils. Aging of colloids is yet one more process leading to the increase of resistance to rupture. Addition of organic matter as sludge lead to the reduction in cohesion (Chang et al., 1983). Overall, dry consistency provides information about soil constituents that is additional to the plasticity class.

Using structural and consistence parameters along with textural classes creates broad subgroups of textural classes within which structural and consistence parameters provide further subdivision (Fig. 6). This implies that the aggregation of 12 textures into five classes (Soil Survey Division Staff, 1993) may be sufficient for the adequate estimation of water retention. Using this aggregation as input for estimating water retention could also help to mitigate field misclassification of textures and presents an interesting avenue to explore.

Although regression tree techniques has provided an interpretable partitioning and has shown internal relationships for the database in this work, it has limitations that preclude addressing several issues that might be of interest in some studies. Other multidimensional classification techniques should be used to find out whether a holistic representation of soil structure with the triplet of size, grade, and shape categories may have more predictive power as compared with using each of those structural parameters as independent predictors. Errors in the regression tree estimates include the effect of field misclassification errors in categories if texture, structure, and consistence. To decompose regression errors and separate effects of misclassification, other regression techniques, such as *dummy coding* (McCullagh and Nelder, 1989), should be used provided the misclassification errors are known. A version of the dummy coding was successfully used by Lin et al. (1999b) to estimate K_{sat} from morphometric indices. Categories of texture, structure, and consistence are not uncorrelated (Fig. 7). That does not preclude using them as independent predictors in regression. Correlation between predictors, or multicollinearity, is a common phenomenon in regression analysis. Neter and Wasserman (1974, p. 341) discuss the milticollinearity in great detail and note that the fact that some of all independent variables are correlated among themselves does not, in general, inhibit an ability to obtain a good fit, nor it tends to affect inferences about mean responses or predictions of new observations, provided these inferences are made within the region of observations. Dropping one or several independent variables from the model will not help to assess the effects of the independent variables for two reasons. First, no information is obtained about dropped independent variables. Second, the magnitudes of the regression coefficients for the independent variables remaining in the model are affected by the correlated independent variables not included in the model (Neter and Wasserman, 1974, p. 346). Nevertheless, should a physics-based method to remedy the multicollinearity in texture, structure, and consistence data be proposed, the dummy coding could be used to apply classical linear statistical regression analysis.

We stress that the results of the regression tree application in this work do not imply any causal relationship between soil consistence and soil structural parameters on one hand, and water retention on another hand. Regression tree merely reflects the fact that structural parameters, consistence parameters, and water retention are affected by the same basic soil properties, that is, content and type of clay minerals, organic matter content and quality, etc. Because those relationships do not show strong correlations, and consistence and structural parameters are not uncorrelated, the accuracy of regression tree predictions is not high, although it may be sufficient for some applications. However, we found it remarkable that qualitative observations of soil morphology and responses to various forms of stress can be translated in quantitative soil hydraulic parameters. Both consistence and structural parameters are useful predictors of soil water retention.

CONCLUSION

The National Soil Characterization database presents a substantial amount of data to search for relationships between both soil structural and consistence parameters on one hand, and soil water retention from the other hand. Regression trees provide a tool to translate the categorical information about soil structure and consistence into quantitative characteristics of water retention. Partitioning the database by soil plasticity class, grade class, and dry consistency class, augmented in some cases by partitioning by the size and shape of structural units, forms mutually exclusive groups of samples. Average value of water retention for each group serves as an estimate for the whole group. Grouping is different for −33 and −1500 kPa water retention. Increase in plasticity, stronger grade for nonplastic soils, and harder dry consistency lead to the increase in water retention. Using soil structural and consistence parameters along with textural classes provides a small, although significant improvement in accuracy of water retention estimates, as compared with estimation from texture alone. Soil structural and consistence parameters can serve as predictors of soil water retention because those parameters reflect soil basic properties that affect soil hydraulic properties.

ACKNOWLEDGMENTS

This work was partially supported by the National Aeronautics and Space Administration Land Surface Hydrology Program.

REFERENCES

Anderson, J.L., and J. Bouma. 1973. Relationships between saturated hydraulic conductivity and morphometric data of an argillic horizon. Soil Sci. Soc. Am. Proc. 37:408–413.

Anderson, D.L., K.M. Portier, T.A. Obreza, M.E.Collins, and D.J. Pitts. 1999. Tree regression analysis to determine effects of soil variability on sugarcane yields. Soil Sci. Soc. Am. J. 63:592–600.

Baker, F.A. 1993. Classification and regression tree analysis for assessing hazard of pine mortality caused by *Heterobasidion annosum*. Plant Dis. 77:136–139.

Bell, J.F. 1999. Tree-based methods. p. 84–111. *In* A.H. Fielding (ed.) Machine learning methods for ecological applications. Kluwer Academic Publ., Boston.

Bicki, T.J., T.E. Fenton, H.D. Luce, and T.A. Dewitt. 1988. Comparison of percolation test results and estimated hydraulic conductivities for Mollisols and Alfisols. Soil Sci Soc Am J. 52:1708–1714.

Bouma, J. 1992. Effect of soil structure, tillage, and aggregation upon soil hydraulic properties. p. 1–36. *In* R.J. Wagenet et al. (ed.) Interacting processes in soil science. Lewis Publ., Boca Raton, FL.

Bouma, J., and J.L. Anderson. 1973. Relationships between soil structure characteristics and hydraulic conductivity. p. 77–105. *In* R.R. Bruce (ed.) Field Soil Water Regime. SSSA Spec. Publ. 5. SSSA, Madison, WI.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1993. Regression trees. Chapman & Hall and CRC Press, Boca Raton, FL.

Carter, M., and S.P. Bentley. 1991. Correlation of soil properties. Pentech Press, London.

Chancellor, W.J. 1994. Friction between soil and equipment materials: a review. Paper no. 94-1002/94-1034. American Society of Agricultural Engineers, St. Joseph, MI.

Chang, A.C., A.L. Page, and J.E. Varneke. 1983. Soil conditioning effects of municipal sludge compost. J. Environ. Eng. 109:574–583.

Clark, L.A., and D. Pregibon. 1992. Tree-based models. p. 377–419. *In* T.J. Hastie (ed.) Statistical models in S. Wadsworth, Pacific Grove, CA.

Danalatos, N.G., C.S. Kosmas, P.M. Driessen, and N. Yassoglou. 1994. Estimation of the draining soil moisture characteristics from standard data as recorded in soil surveys. Geoderma 64:155–165.

Domzal, H. 1970. Preliminary studies of the influence of moisture on physico-mechanical properties of some soils with regard to estimation of optimum working conditions of implements. Polish J. Soil Sci. 3:61–70.

Fielding, S.H. (ed.) 1999. Machine learning methods for ecological applications. Kluwer Academic Publ., Dordrecht, The Netherlands.

Filgueira, R.R., Ya. Pachepsky, L.L. Fournier, G.O. Sarli, and A. Aragon. 1999. Comparison of fractal dimensions estimated from aggregate mass-size distribution and water retention scaling. Soil Sci. 164:217–226.

Gimenez, D., W.J. Rawls, Y. Pachepsky, and J.P.C. Watt. 2001. Prediction of a pore distribution factor from soil textural and mechanical parameters. Soil Sci. 166:79–88.

Good, P.I. 1999. Resampling methods: a practical guide to data analysis. Birkhäuser, Boston.

Griffiths, E. 1991. Assessing permeability class from soil morphology. Tech. Record no. 40. New Zealand DSIR Land Resources, Auckland, New Zealand.

Hall, D.G.M., M.J. Reeve, A.I. Thomasson, and V.F. Wright. 1977. Water retention, porosity and density of field soils. Soil Surv. Tech. Monogr. 9. Harpenden:Rothamsted Exp. Stn., UK.

Horn, R. 1993. Mechanical properties of structured unsaturated soils. Soil Technol. 6:47–75.

Horn, R., and T. Baumgartl. 1999. Dynamic properties of soils. p. A19–A50. *In* M.E. Sumner (ed.) Handbook of soil science. CRC Press, Boca Raton, FL.

Ibanga, I.J., O.W. Bridwell, W.L. Powers, A.M. Feyerherm, and W.W. Williams. 1980. Soil consistence: effect of particle size. Soil Sci. Soc. Am. J. 44:1124–1126.

Van Lanen, H.A.J., C.A.J. van Diepen, G.J. Reinds, and G.H.J. de Koning. 1992. A comparison of qualitative and quantitative physical land evaluations, using an assessment of the potential for sugarbeet growth in the European Community. Soil Use Manage. 8:80–89.

Leenhardt, D. 1995. Errors in estimation of soil water properties and their propagation through a hydrological model. Soil Use Manage. 11:15–21.

Lees, B.G., and K. Ritman. 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environ. Manage. 15:823–831.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999a. Effects of soil morphology on hydraulic properties. I. Quantification of soil morphology. Soil Sci. Soc. Am. J. 63:948–954.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999b. Effects of soil morphology on hydraulic properties: II. Hydraulic pedotransfer functions. Soil Sci. Soc. Am. J. 63:955–961.

Lysenko, M.P. 1972. Composition and physical-mechanical properties of soils. (In Russian.) Nedra, Moscow.

Mathsoft. 1999. SPLUS 2000 Professional—User's Manual. Mathsoft, Seattle, WA.

McCullagh, P., and J.A. Nelder. 1989. Generalized linear models. 2nd ed. Chapman & Hall, London.

McKeague, I.A. 1987. Estimating air porosity and available water capacity from soil morphology. Soil Sci. Soc. Am. J. 51:148–152.

McKenzie, N.J., and D.W. Jacquier. 1997. Improving the field estimation of saturated hydraulic conductivity in soil survey. Aust. J. Soil Res. 35:803–825.

McKenzie, N.J., and P.J. Ryan. 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94.

McKenzie, N.J., K.R.J. Smettem, and A.J. Ringrose-Voase. 1991. Evaluation of methods for inferring air and water properties of soils from field morphology. Aust. J. Soil Res. 29:587–602.

Michaelsen, J., D.S. Schimel, M.A. Friedl, F.W. Davis, and R.C. Dubayah. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J. Vegetation Sci. 5:673–686.

Neter, J., and W. Wasserman. 1974. Applied linear statistical models. Regression, analysis of variance, and experimental designs. Richard D. Irwin Publ., Homewood, IL.

Nettleton, W.D., C.S. Hozney, K.W. Flach, and B.R. Brasner. 1969. Soil scientists' proficiency in describing soil consistence. Soil. Sci. Soc. Am. Proc. 33:320–321.

Nimmo, J.R. 1997. Modeling structural influences on soil water retention. Soil Sci. Soc. Am. J. 61:712–719.

Pachepsky, Ya.A., W. Rawls, D. Gimenez, and J.P.C. Watt. 1998. Use of soil penetration resistance and group method of data handling to improve soil water retention estimates. Soil Tillage Res. 49:117–128.

Post, D.F., A.R. Huete, and D.S. Pease. 1986. A comparison of soil scientist estimations and laboratory determinations of some Arizona soil properties. J. Soil Water Conserv. 41:421–424.

Rawls, W.J., D.L. Brakensiek, and S.D. Logsdon. 1993. Predicting saturated hydraulic conductivity utilizing fractal principles. Soil Sci. Soc. Am. J. 57:1193–1197.

Rawls, W.J., T.J. Gish, and D.L. Brakensiek. 1991. Estimating soil water retention from soil physical properties and characteristics. Adv. Soil Sci. 16:213–234.

Schaap, M.G., and F.J. Leij. 1998. Database-related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163:765–779.

Shaw, J.N., O.T. West, C.C. Truman, and D.E. Radcliffe. 1997. Morphologic and hydraulic properties of soils with water restrictive horizons in the Georgia Coastal Plain. Soil Sci. 162:875–885.

Singh, K.K., T.S. Colvin, D.C. Erbach, and A.Q. Mughal. 1992. Tilth index: an approach to quantifying soil tilth. Trans. ASAE 35:1777–1785.

Soil Survey Division Staff. 1993. Soil survey manual. USDA Handbook 18. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff. 1997. National characterization data. Soil Survey Laboratory, National Soil Survey Center, and Natural Resources Conservation Service, Lincoln, NE.

Southard, R.J., and S.W. Buol. 1988. Subsoil saturated hydraulic conductivity in relation to soil properties in North Carolina Coastal Plain. Soil Sci. Soc. Am. J. 52:1091–1094.

Vepraskas, M.J., W.R. Guertal, H.J. Kleiss, and A. Amoozegar. 1996. Porosity factors that control the hydraulic conductivity of soil-saprolite transitional zones. Soil Sci. Soc. Am. J. 60:192–199.

Voronin, A.D. 1990. Energy-based concept of soil physical status. Pochvovedenie 5:7–19.

Wang, C., J.A. McKeague, and G.C. Topp. 1985. Comparison of estimated and measured horizontal K_{sat} values. Can. J. Soil Sci. 65: 707–715.

Williams, J.R., E. Prebble, W.T. Williams, and C.T. Hignett. 1983. The influence of texture, structure and clay mineralogy on the soil moisture characteristic. Aust. J. Soil Res. 21:15–31.

Williams, J., P. Ross, and K. Bristow. 1992. Prediction of the Campbell water retention function from texture, structure, and organic matter. p. 427–442. *In* M.Th. van Genuchten et al. (ed.) Proc. Int. Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils, Riverside, CA. 11–13 Oct. 1989. Univ. of California, Riverside, CA.